DOCUMENT RESUME

ED 379 294 TM 022 639

AUTHOR Gearhart, Maryl
TITLE Toward the Instructional Utility of Large-Scale Writing Assessment: Validation of a New Narrative Rubric. Project 3.1. Studies in Improving Classroom and Local Assessments. Portfolio Assessment: Reliability of Teachers' Judgments.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Jul 94
CONTRACT R117G10027
NOTE 50p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Educational Assessment; Elementary Education; Holistic Approach; *Interrater Reliability; *Narration; *Scoring; Student Evaluation; Test Construction; *Validity; Writing Evaluation
IDENTIFIERS Apple Classrooms of Tomorrow; *Large Scale Writing Assessment; *Writing What You Read

ABSTRACT

The "Writing What You Read" (WWYR) rubric was designed for large-scale assessments, and differs from most narrative rubrics in its narrative-specific content and its developmental framework. The rubric contains five analytic subscales for theme, character, setting, plot, and communication, and a sixth holistic scale for overall effectiveness. Evidence of validity was gathered for the WWYR scoring rubric through comparison with an established narrative rubric that has been demonstrated to be sound. The comparison rubric, derived from comparative studies of student writing competence, is a holistic/analytic scheme used annually in California. Five raters reviewed narrative samples collected from an elementary school participating in the Apple Classrooms of Tomorrow project. Both rubrics were generally used consistently by raters. Results suggest that at least three subscales of WWYR can be used reliably and meaningfully in large-scale assessment as long as each narrative is rated by two raters. Evidence is lacking for the technical soundness of the other scales, and findings further suggest that subscale judgments may not provide a technically sound profile of students' strengths and weaknesses. One figure and 12 tables present study findings. (Contains 23 references.) (SLD)

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – July 1994

Project 3.1 Studies in Improving Classroom
and Local Assessments

Portfolio Assessment: Reliability of Teachers' Judgments

Toward the Instructional Utility of Large-Scale Writing Assessment: Validation of a New Narrative Rubric

Maryl Gearhart, Project Study Director

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported in part under the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

Acknowledgments

Our five raters, Kathy Beneiof, Virginia Espinosa, Rosa Valdes, Maria Vega-Vidales, and Kim Uebelhardt, made important and substantial contributions that we summarize in detail in the paper. Jia Moody contributed statistical advice and assistance. Edys Quellmalz, Andrea Whittaker, and Stevens Creek teachers provided helpful input to the design of the *Writing What You Read* rubric. Our thanks to the teachers, students, and parents who permitted us access to the students' narratives.

TOWARD THE INSTRUCTIONAL UTILITY OF LARGE-SCALE WRITING ASSESSMENT: VALIDATION OF A NEW NARRATIVE RUBRIC

Maryl Gearhart,[1] Joan L. Herman,[1] John R. Novak,[1]
Shelby A. Wolf,[2] and Jamal Abedi[1]

In the press to design performance-based writing assessments to serve both policy and practice, scoring rubrics have undergone considerable scrutiny and revision (Freedman, 1991; Huot, 1991; Paul, 1993; Wiggins, 1993). While concerns for large-scale assessment and policy uses have emphasized requirements for technical quality—particularly the capacity to support interrater agreement—, interest in instructional value and impact on practice have highlighted the importance of rubric content and structure.

Two related issues have emerged in the content dialogue. First, existing rubrics cannot adequately represent the important qualities of good writing when scales or scale-point criteria are vague, confusing, or inconsistent with what is known about well-constructed and effective text (Baxter, Glaser, & Raghavan, in press; Paul, 1993; Resnick, Resnick, & DeStefano, 1993; Wiggins, 1993; Wolf, 1993).

> Most of the scoring rubrics that I have encountered seem invalid to me. [W]e score what is easy, uncontroversial, and typical—not necessarily what is apt for identifying exemplary writing or apt for the demands of real-world writing. Consider [one state's] . . . descriptor for the top score on the scale [of] Organization/Content . . . Little in this scoring system places a premium on style, imagination, or ability to keep the reader interested. Only the top score description mentions "effective and vivid" responses, instead of those criteria being woven through the whole rubric. Yet we see this limitation in almost every writing assessment. (Wiggins, 1993, p. 21)

Second, rubrics that do not reflect the qualities of good writing are limited in their instructional utility (Paul, 1993; Wiggins, 1993). If a central purpose of assessment is to guide instructional planning, then rubrics for assessing student writing must be derived from current English/language arts frameworks and must reflect those analyses of the contents, purposes, and

[1] CRESST/University of California, Los Angeles. [2] CRESST/University of Colorado at Boulder.

complexities of text. Rubrics must communicate to teachers, students, and others what's important in writing performance.

Certainly the challenges to rubric design are substantial. The purpose of the study reported here was to validate a new rubric designed to optimize content quality and to enhance instructional value, but whose technical quality is unknown. The design of the *Writing What You Read* (WWYR) narrative rubric (Wolf & Gearhart, 1993a, 1993b) was prompted by the need for judgments that "chart . . . the course between uniformity of judgment on the one hand and representation of complexity and diversity on the other hand" (Wolf, Bixby, Glenn, & Gardner, 1991). That need is particularly crucial for classroom teachers who are concerned not only with students' present work, but with their future growth. Existing narrative rubrics did not, in our judgment, have the potential to guide instruction.

> [F]or example, . . . in a pilot project to score locally completed work . . . using the [a national] rubric, . . . [h]ere is a descriptor for a story that merits a score of 6 (the top level): "Paper describes a sequence of episodes in which almost all story elements are well developed (i.e., setting, episodes, characters' goals, or problems to be solved). The resolution of the goals or problems at the end are [sic] elaborated. The events are represented and elaborated in a cohesive way." Surely this is not the best description possible of a good story. (Wiggins, 1993, p. 21)

Surely not. But could the "best description," or even a better description, be captured in a technically sound rubric? Our recognition of the "test-maker's dilemma" (Wiggins, 1993)—that rubric complexity and face validity could result in a loss of technical capacity for large-scale use--was the impetus for the study reported here.

Issues in the Design of Writing Rubrics

A first decision involved type of rubric. Three types of scoring rubrics have been used in large-scale writing assessment. First, and probably most common, is holistic scoring—assignment of a single score reflecting a student's competence with all aspects of writing. A second approach is primary trait scoring—the construction of a rubric customized to specific prompts. A third method is the analytic rubric, in which defined dimensions of good writing (e.g., Organization, Content, Style, Voice, and Mechanics) are applied across a range of topics within broadly defined genres.

Narrative Rubric

Theme

explicit ◄——► implicit
didactic ◄——► revealing

- Not present or not developed through other narrative elements
- Meaning centered in a series of list-like statements ("I like my mom And I like my dad. And I like my. ") or in the coherence of the action itself ("He blew up the plane Pow!")
- Beginning statement of theme—often explicit and didactic ("The mean witch chased the children and she shouldn't have done that "); occasionally the theme, though well stated, does not fit the story
- Beginning revelation of theme on both explicit and implicit levels through the more subtle things characters say and do ("He put his arm around the dog and held him close 'You're my best pal,' he whispered ")
- Beginning use of secondary themes, often tied to overarching theme, but sometimes tangential, main theme increasingly revealed through discovery rather than delivery, though explicit thematic statements still predominate
- Overarching theme multi-layered and complex; secondary themes integrally related to primary theme or themes, both explicit and implicit revelations of theme work in harmony ("You can't do that to my sister!", Lou cried, moving to shield Tasha with her body)

Character

flat ◄——► round
static ◄——► dynamic

- One or two flat, static characters with little relationship between characters; either objective (action speaks for itself) or first person (author as "I") point of view
- Some rounding, usually in physical description; relationship between characters is action-driven, objective point of view is common
- Continued rounding in physical description, particularly stereotypical features ("wart on the end of her nose"); beginning rounding in feeling, often through straightforward vocabulary ("She was sad, glad, mad.")
- Beginning insights into the motivation and intention that drives the feeling and the action of main characters often through limited omniscient point of view; beginning dynamic features (of change and growth)
- Further rounding (in feeling and motivation), dynamic features appear in the central characters and in the relationships between characters; move to omniscient point of view (getting into the minds of characters)
- Round, dynamic major characters through rich description of affect, intention, and motivation; growth occurs as a result of complex interactions between characters, most characters contribute to the development of the narrative; purposeful choice of point of view

Setting

backdrop ◄——► essential
simple ◄——► multi-functional

- Backdrop setting with little or no indication of time and place ("There was a little girl. She liked candy.")
- Skeletal indication of time and place often held in past time ("once there was..."); little relationship to other narrative elements
- Beginning relationship between setting and other narrative elements (futuristic setting to accommodate aliens and spaceships); beginning symbolic functions of setting (often stereotypical images—forest as scary place)
- Setting becomes more essential to the development of the story in explicit ways: characters may remark on the setting or the time and place may be integral to the plot
- Setting may serve more than one function and the relationship between functions is more implicit and symbolic—for example, setting may be linked symbolically to character mood ("She hid in the grass, clutching the sharp, dry spikes, waiting.")
- Setting fully integrated with the characters, action, and theme of the story; role of setting is multifunctional—setting mood, revealing character and conflict, serving as metaphor

Plot

simple ◄——► complex
static ◄——► conflict

- One or two events with little or no conflict ("Once there was a cat. The cat liked milk.")
- Beginning sequence of events, but occasional out-of-sync occurrences; events without problem, problem without resolution, or little emotional response
- Single, linear episode with clear beginning, middle, and end; the episode contains four critical elements of problem, emotional response, action, and outcome
- Plot increases in complexity with more than one episode; each episode contains problem, emotional response, action, outcome; beginning relationship between episodes
- Stronger relationship between episodes (with the resolution in one leading to a problem in the next); beginning manipulation of the sequence through foreshadowing, and subplots
- Overarching problem and resolution supported by multiple, episodes; rich variety of techniques (building suspense, foreshadowing flashbacks, denouement) to manipulate sequence

Communication

context-bound ◄——► reader-considerate
literal ◄——► symbolic

- Writing bound to context (You have to be there) and often dependent on drawing and talk to clarify the meaning; minimal style and tone
- Beginning awareness of reader considerations; straightforward style and tone focused on getting the information out; first attempts at dialogue begin
- Writer begins to make use of explanations and transitions ("because" and "so"); literal style centers on description ("sunny day"); tone explicit
- Increased information and explanation for the reader (linking ideas as well as episodes); words more carefully selected to suit the narrative's purpose (particularly through increased use of detail in imagery)
- Some experimentation with symbolism (particularly figurative language) which shows reader considerations on both explicit and implicit levels; style shows increasing variety (alliteration, word play, rhythm, etc.) and tone is more implicit
- Careful crafting of choices in story structure as well as vocabulary demonstrate considerate orchestration of all the available resources; judicious experimentation with variety of stylistic forms which are often symbolic in nature and illuminate the other narrative elements

Figure 1. Narrative rubric.

Wolf/Gearhart/Quellmalz/Whittaker, 1993

Advocates of these scoring approaches debate their efficiency, cost effectiveness, and relative value for instructional feedback. Although empirical comparisons frequently show significant correlations among analytic subscale scores and between holistic and analytic scoring, it is our view that concerns about instructional utility press for feedback beyond a single score. The scores produced by holistic scoring do not communicate the far more complex standards articulated by raters in moderation sessions, and therefore holistic scores are of limited value to the recipients. In contrast, analytic scales have greater instructional potential, in that they communicate a differentiated analysis of quality standards. They can do so, however, only *if* the dimensions reflect consensus on the components of valued performances. The design of the *Writing What You Read* rubric was motivated by the concern to ground rubric design in current analyses of effective narrative writing.

The *Writing What You Read* Rubric

The *Writing What You Read* rubric (Figure 1) differs from most narrative rubrics in its narrative-specific content and its developmental framework (Wolf & Gearhart, 1993a, 1993b). Designed for classroom use, the rubric contains five analytic subscales for Theme, Character, Setting, Plot, and Communication (Figure 1), and a sixth, holistic scale for Overall Effectiveness constructed specifically for this technical study (Table 1). Each subscale contains six levels designed to match current understandings of children's narrative development. The rubric was the product of collaboration with elementary teachers, and its use has been shown to impact teachers' understandings of narrative (Gearhart, Wolf, Burkey, & Whittaker, 1994). It has never been used to date for large-scale assessment.

The technical language of narrative is integral to the WWYR tool, unlike the descriptors of many narrative rubrics that are not unique to this genre. Words like topic (rather than theme), event (rather than episode), and diction (rather than style) create a sense of "genre generality" (Gearhart et al., 1994). When narrative components are included, they are usually limited to character, setting, and plot, omitting theme—the heart of narrative, a comment about life which illuminates the emotional content of the human condition. A subscale for organization may not capture the orchestration of components. Definitions for the narrative's development may omit the communicative aspects of style and tone, focusing instead on logical

Table 1

Writing What You Read: Overall Effectiveness—How are features integrated in this narrative?

1. A character suspended without time, place, action, or conflict. More a statement than a narrative.	*There was a little girl who liked rainbows.* *Poor little Cyclops. He had one eye.*
2. Action-driven narrative written in list-like statements. Character(s) and setting minimal. Plot minimal or missing key pieces in sequence, conflict, or resolution.	*Sleeping Beauty has a prince. She had a balloon and a kite. The sun was very beautiful and shining. She went to a party and she had fun. She had a party dress on and her prince.* *Once there was a little girl . And she was 10 years old. And she was very beautiful. A big bear came out of the forest and she ran deep in the forest. Her name is Amelia. But he was going for Amelia. The little girl was very scared. But then she was happy.*
3. One episode narrative (either brief or more extended) which includes the four critical elements of problem, emotional response, action, and outcome. One or more of these elements may be skeletal. The characters and setting are related but often fairly stereotypical, as is the language which describes them.	See *The Dragon Fight* and *The True Three Little Pigs* in the Guidebook. A fable would fit here. *One there was a little girl. Her name was Ashley. She was very pretty. She had red hair and freckles. She also had beautiful brown eyes like brown lakes. Anyway...she was a princess that lived in a golden castle. Her father was the king of the land.* *Oh! I forgot! Ashley had a big sister that was not mean. Her name was Lindsey. And she was just as beautiful as Ashley, but she had brown hair.* *Now the real problem was the grandma. She did not like the children. She thought they were spoiled brats. But the children loved their grandmother.* *It so happened that the grandmother had made a plan so the next day the children would die. And this is how it turns out.* *Well, you see, this woman was not the ordinary grandmother. She actually was a witch. Anyway, she decided to have them go and take a walk in the forest. Then she put a pretty flower out in the path. She knew they would notice it. (If you touched the flower and then touched your hair without washing your hair before two day's time you would die!)* *The next day the girls took a walk in the forest and everything was going as the witch had planned except a couple of drops of water landed in the place where the flower had touched the children's hair.* *When the children came home, the grandma was so angry to see them alive that she jumped off a cliff and was never seen again*

Table 1 (continued)

4. More than one episode narrative with greater insight into character motivation. Beginning revelation of theme on double levels (both implicit and explicit), and setting is more essential to the tale. Language more detailed, more suited to the narrative, and offers careful transitions.	See *The Seven Chinese Brothers* (from the youngest's point of view) in the Guidebook. Examples from the story appear under Character and Communication. *The True Story of Cinderella — Dedicated to all the badly treated, beautiful maidens of the world. And the beautiful Fairy godmothers that help them.* *Once upon a time, long ago and far away, there lived Cinderella, and her two ugly step-sisters and one step-mother. They lived in Hollywood in the biggest castle ever made and of all people Cinderella was the poor little servant.* *One night Cinderella had more work than usual. She had to sew dresses and put make-up on her two step-sisters and her ugly mean step-mother. They were going to the prince's ball. The prince was to find a wife. When her step-sisters and step-mother left Cinderella, she started to cry. She wanted to go with her step-mother and step-sisters. All of a sudden a big puff of smoke filled the air and here I am.* *I said that I was her fairy god mother. I am going to help her go to the ball and dance with the prince for the whole night. But as Cinderella turned her head I saw how desperate she really was. But I felt that a man just wants someone to do their dishes and their dirty work for them. Still, she was deeply in love.* *This was where the magic comes in. I took the apple from the table and waved my magic wand above my head and the apple turned into a magical carriage. I took my magic wand and waved it over Cinderella's head and said, "Turn this filthy little maid into a beautiful princess."* *I took the ants off the other fruit and turned them into horses for the ride there. I looked at her. She was the most beautiful woman I ever saw. Then Cinderella asked, "Why didn't you come before?"* *"I was busy babysitting Goldilocks."* *Then Cinderella and I stepped into the carriage, and we rode into the night. On the way there I told her that she would have to be back by midnight, or the magic will wear out, and she would be the same dirty little girl that she was before. When they got there I changed her ugly step-sisters and step-mother into frogs. Cinderella danced with the prince for the rest of the night. The next day they got married. They lived happily ever after.*

Table 1 (continued)

5. Multilayered narrative with connected episodes. Character and setting description are detailed and sometimes symbolic to reveal intention, motivation, and integration of individuals with time and space. There is evidence of some risk-taking in plot manipulation (e.g., efforts to foreshadow or embed subplots) and experimentation with language (e.g., figurative language, word play).	*Once there was a king and queen who lived in a golden castle of great beauty, but they had no children. Finally, they had a daughter. They had a splendid feast and they invited all the fairies to court except the eldest fairy because she was a wicked witch.* *When it was time to give the wishes, the eldest fairy stormed in and said, "I curse the child! " Her voice sounded like stones falling from a cliff. "She shall be ugly and when she is fifteen she shall look into a mirror and die!"* *After the wicked witch left, the youngest fairy said, "She shall not die, but just faint for 100 years. However, I cannot change the ugliness. My little wand cannot overpower the eldest fairy." So the king broke all the mirrors in the castle.* *As the ugly princess grew up, it was very hard because everybody in the court teased her. Yet, the servants in the castle loved her as they would their own daughter.* *Time went by and the ugly princess turned fifteen and she decided that she would explore the castle. She went into a tower and there she saw an old woman putting clips into her hair while staring into an odd square of glass that reflected the old woman's face.* *The ugly princess said, "May I try?" She took a clip, and when she stepped before the mirror, she saw her horrible face and fell in a faint to the floor. The witch laughed and said, "I've got you now!"* *Soon, however, the little fairy came and picked up the princess and laid her on a little bed where she slept for a hundred years. But the wicked witch's magic was so powerful that everyone in the castle fell asleep too.* *At the end of the hundred years, an unattractive prince was riding by on a disgusting-looking horse, when he chanced to see a torn up flag fluttering from the tip of a distant tower.* *Then he stopped and remembered a story he had heard when he was only a boy about an ugly princess. Since he hadn't had any luck with beautiful princesses during his journey, he decided to try an ugly one.* *He went into the quiet castle. His footsteps echoed in the halls. Nothing stirred. He felt like the walls were holding their breath. Then he saw a tiny stairway and climbed it to the tower room. When he entered the room, he saw the Sleeping Ugly. He bent to kiss her, but then he stopped and said, "Should I be doing this." But then he decided even though she was ugly on the outside, she was probably very beautiful on the inside.* *He kissed her and she woke up. They were married in a beautiful green meadow with daisies all around. They had two ugly children and they lived happily ever after in a castle without mirrors for the rest of their lives.*
6. A rich and multilayered narrative with fully integrated, often multifunctional components, and considerable orchestration in communication to illuminate the components. Growth in characters, purposeful point of view, variety of plot techniques, crafted choice of language.	No example available.

transitions; although transitions are important and logic is always welcome, the communicative aspects of narrative are more centered on creating images—using language purposefully, metaphorically, and rhythmically to take the reader off the page and into another world.

WWYR was designed as an alternative to narrative rubrics that are not grounded in *genre*, either in its traditional sense of a classification system for organizing literature (a system much subject to change) or in its more current sense of social action constrained by particular rhetorical forms. The development of character, the symbolism in setting, the complexity of plot, the subtlety of theme, the selected point of view, and the elaborate use of language all depend on and are defined by genre. If we are going to teach children about narrative and how to grow as young story writers, then surely we would want to use more precise language and to provide a fuller picture of what narrative is. If we limit or simplify concepts for children (and for their teachers), we refuse them access to more intriguing and more authentic possibilities. The WWYR rubric is a simplification, of course—how else could something as complex as narrative fit on a single page? Yet, its language and focus provide a key to a much larger door, opening onto the evocative, emotional, and eminently human symbol system of narrative meaning.

Our Study

The purpose of our study was to gather evidence of validity for the *Writing What You Read* rubric, through technical comparisons with an established narrative rubric that has consistently demonstrated sound technical capabilities in large-scale assessments of elementary level writing. Our studies addressed a series of questions regarding the technical quality of the rubric.

> *Reliability: Can the* Writing What You Read *rubric be applied to scoring of classroom narratives with the same levels of rater agreement as an established narrative rubric?*

We selected a comparison narrative rubric that has consistently demonstrated excellent levels of rater agreement. Can raters make judgments with the *Writing What You Read* rubric at similar levels of reliability? To investigate

this question, judges rated classroom narratives with both rubrics, and we compared reliabilities.

Validity of the Writing What You Read *rubric: What is the evidence that scores derived from the WWYR rubric are meaningful indices of students' narrative writing?*

We inferred validity from grade-level differences (scores should increase with age), from relationships of scores across types of assessments (e.g., scores derived from both rubrics should be correlated), from interscale correlations (for both rubrics, the subscale ratings should not be highly intercorrelated), from consistency of raters' judgments across rubrics, and from raters' confidence in their judgments based on opinions expressed in post-rating interviews.

Procedures

Site

The narrative samples were collected from an elementary school that served as a longitudinal research site for the national Apple Classrooms of Tomorrow (ACOT℠) project. The school is located in a middle class suburb of Silicon Valley.

Datasets

Narratives were sampled from classroom writing in Grades 1 through 6. Students' names and grade levels were removed and replaced with identification numbers. Narratives were sorted by level (primary = Grades 1 and 2, middle = Grades 3 and 4, and upper = Grades 5 and 6), and then scrambled within sets.

Comparison Rubric

The comparison rubric, derived from analytic scales used in the IEA comparative studies of student writing competence, is a holistic/analytic scheme (Table 2). (See Quellmalz & Burry, 1983, for description of original CSE scales.) In annual use in assessments of students' narratives in a California school district, this rubric has also been used extensively in our Center for evaluations of elementary students' writing (Baker, Gearhart, & Herman, 1990, 1991; Baker, Herman, & Gearhart, 1988; Gearhart, Bank, & Herman,

Table 2

Comparison Narrative Rubric

General Competence	Focus/Organization	Development	Mechanics
6 EXCEPTIONAL ACHIEVEMENT EXCEPTIONAL WRITER	- topic clear - events logical - no digressions - varied transitions - transitions smooth and logical - clear sense of beginning and end	- elements of narrative are well-elaborated (plot, setting, characters) - elaboration even and appropriate - sentence patterns varied and complex - diction appropriate - detail vivid and specific	- one or two minor errors - no major errors
5 COMMENDABLE ACHIEVEMENT COMMENDABLE WRITER	- topic clear - events logical - possible slight digression without significant distraction to reader - most transitions smooth and logical - clear sense of beginning and end	- elements of narrative are well-elaborated - most elaboration is even and appropriate - some varied sentence pattern used - vocabulary appropriate - some details are more vivid or specific than general statements - a few details may lack specificity	- a few minor errors - one or two major errors - no more than 5 combined errors (major and minor) - errors do not cause significant reader confusion
4 ADEQUATE ACHIEVEMENT COMPETENT WRITER	- topic clear - most events are logical - some digression causing slight reader confusion - most transitions are logical but may be repetitive - clear sense of beginning and end	- most elements of narrative are present - some elaboration may be less even and lack depth - some details are vivid or specific although one or two may lack direct relevance - supporting details begin to be more specific than general statements	- a few minor errors - one or two major errors - no more than 5 combined errors (major and minor) - errors do not cause significant reader confusion

13

Table 2 (continued)

General Competence	Focus/Organization	Development	Mechanics
3 SOME EVIDENCE OF ACHIEVEMENT DEVELOPING WRITER	- topic clear - most events logical - some digression or over-elaboration interfering with reader understanding - transitions begin to be used - limited sense of beginning and end	- elements of narrative are not evenly developed, some may be omitted - vocabulary not appropriate at times - some supporting detail may be present	- some minor errors - some major errors - no fewer than 5 combined errors (major and minor) - some errors cause reader confusion
2 LIMITED EVIDENCE OF ACHIEVEMENT EMERGING WRITER	- topic may not be clear - few events are logical - may be no attempt to limit topic - much digression or overelaboration with significant interference with reader understanding - few transitions - little sense of beginning or end	- minimal development of elements of narrative - minimal or no detail - detail used is uneven and unclear - simple sentence patterns - very simplistic vocabulary - detail may be irrelevant or confusing	- many minor errors - many major errors - many errors cause reader confusion and interference with understanding
1 MINIMAL EVIDENCE OF ACHIEVEMENT INSUFFICIENT WRITER	- topic is clear - no clear organizational plan - no attempt to limit topic - much of the paper may be a digression or elaboration - few or no transitions - almost no sense of beginning and end	- no development of narrative elements - no details - incomplete sentence patterns	- many major and minor errors causing reader confusion - difficult to read

1990; Gearhart, Herman, Baker, & Whittaker, 1992; Gearhart, Herman, & Bank, 1989; Herman, Gearhart, & Baker, 1994). Consistently demonstrating excellent levels of rater agreement and meaningful relationships with indices of instructional emphasis, the rubric represents a sound technical approach to writing assessment. Four 6-point scales are used for assessment of General Competence, Focus/Organization, Elaboration, and Mechanics; in the current study, we were concerned just with narrative content, and the raters did not apply the Mechanics scale.

Rating Procedures

Raters. Our five raters were drawn from three communities. Two raters were elementary teachers with experience using the comparison rubric for scoring students' narrative writing; one of these raters had considerably more experience than the other with district scoring sessions. Two raters were elementary teachers experienced with other large-scale efforts; one scored elementary narrative and persuasive writing samples in English and Spanish for two years as part of a program evaluation, and the other scored writing samples of elementary school students in English and Spanish as part of a nationally implemented supplemental education program. The fifth rater was a research assistant with experience scoring elementary narrative and persuasive writing samples in English and Spanish for program evaluation.

Rating procedures. In conducting the narrative scoring, raters were informed that the samples would represent primary (Grades 1-2), middle (Grades 3-4), or upper (Grades 5-6) elementary levels, and that sets would be labeled by levels. Raters completed comparison scoring before undertaking *Writing What You Read* scoring. While order of rubric is certainly a variable that could impact judgments, we felt that our initial questions regarding the *Writing What You Read* rubric did not require systematic investigation of rubric order at this time.

Each phase of scoring began with study and discussion of each rubric, the collaborative establishment of benchmark papers distributed along the scale points, and the scoring of at least three papers in a row where disagreement among raters on any scale was not greater than 0.5. Raters requested and were granted permission to locate ratings at midpoints in addition to defined scale points. Training papers for each major phase were drawn from all

levels. When raters began the scoring of a given level, they conducted an additional training session; raters scored preselected papers independently, resolved disagreements through discussion, and placed these "benchmark" papers in the center of the table for reference.

Because the set of papers for Grades 3 and 4 was by far the largest, raters rated half of these first, followed by Primary, Upper, and then the remaining Middle papers. Raters revisited the Middle-level benchmark papers when scoring the second half of that set. Raters rated material in bundles labeled with two raters' names; at any given time, each rater made a random choice of a bundle to score. The material was distributed so that two raters rated each piece independently; scores were entered rapidly, and a third rater rated any paper whose scores on any scale differed by more than one scale point. A check set of three to eight papers was included halfway through the scoring session; any disagreements were resolved through discussion that made certain that raters were not changing their criteria for scoring.

Rater Reflection

Raters were interviewed at two key points in the session—at the completion of the comparison scoring, and at the completion of the final *Writing What You Read* scoring. The comparison interview was conducted as a focus group; the final interview was a critique of the two rubrics and was conducted with two pairs of raters and one rater alone. Interviews were transcribed for analysis. The protocols for both interviews are contained in the Appendix.

Results

Rater Agreement

Reliability: Can the Writing What You Read *rubric be applied to scoring of classroom narratives with the same levels of rater agreement as comparison scoring?*

Rater agreement was examined using percent agreement, correlation coefficients, and generalizability coefficients. Because raters utilized midpoint ratings, percent agreement was computed for ± 0, 0.5, and 1.0. Analyses of agreement, correlation coefficients, and generalizability coefficients were

based only on the material rated independently and thus excluded ratings negotiated during the training or the check sets.

Correlation coefficients and percent agreement indices were computed for each pair of raters, and, for purposes of comparison, those estimates were averaged across all pairs of raters. The average percentages of agreement should be considered to be descriptive information rather than evidence of reliability, since given the small range of possible values and the restricted number of scale points, rather high levels of agreement may be expected just based on chance alone. Indeed, repeated estimation of agreement indices after random permutations of the data indicated that, for these scales and these data, the chance levels of agreement for uncorrelated ratings were on the order of .16, .44, and .67 for the ±0, ±0.5, and ±1 indices, respectively. The introduction of very moderate correlations between ratings are sufficient to cause the percentages of adjacent (±1) agreement to approach the ceiling value of 1.00. The average correlations can be interpreted much like classical reliability coefficients, with the difference that instead of estimating the correlation between parallel forms of a test (as in classical reliability theory), we are estimating the correlation between parallel ratings of a single test.

Interrater reliability for both rubrics was also assessed through the use of generalizability theory, a powerful and appropriate methodology for addressing issues of rater agreement. For purposes of discussion here, a generalizability coefficient can be considered to be analogous to the classical reliability coefficient. Both can be computed as ratios of variances. A reliability coefficient is the ratio of an examinee's true score variance to the observed score variance, and it is an estimate of the correlation between scores on parallel forms of a test. Generalizability coefficients are ratios of variance due to the objects of measurement (in our case, students' essay scores) to the total variance due to the objects of measurements and the conditions of measurement (in our case raters). They are estimates of the correlations between observations obtained under different conditions of measurement (by different raters).

Generalizability theory is much more flexible than classical reliability theory in that generalizability coefficients can be tailored to suit the particular purposes of an evaluation. For example, separate generalizability coefficients can be computed for relative and absolute decisions. If one were interested

mainly in accurately ranking a set of essays, then a relative coefficient would be of interest. If relative generalizability is high, then one can be confident that two different raters scoring the same set of essays would create consistent rank orderings of the essays. That is, there would be a high degree of agreement on which essay is the best, which is second best, etc. On the other hand, if one is making decisions about proficiency by comparing scores to an absolute standard, such as a cut score, or is comparing scores assigned by different raters, then an absolute coefficient is more appropriate. This type of coefficient takes into account the variance that is due to differences between raters. If, in the scenario presented above, relative generalizability were high but absolute generalizability were low, then it would be difficult to have confidence in comparisons between means of sets of essays rated by different raters.

Another advantage of generalizability theory is that it is easy to extend the results of a generalizability study (G-study) to what is called a decision study (D-study). In classical test theory, the reliability of the test is a function of the length of the test; longer tests are more reliable, and the reliability of a test can be improved by adding more items. The analogous procedure in a rating situation is to improve reliability by adding more raters, multiply scoring each essay, and aggregating the results. The G-study coefficients in our study can be interpreted as reliability indices for scores based on a single rater. If those coefficients are too low, then a D-study can be done to examine the effects on generalizability of adding more raters. An informed decision can then be made as to how many raters should be used to attain adequate levels of generalizability.

The design for the G-study for this paper utilized essays as the object of measurement, and raters as conditions of measurement. In the parlance of generalizability theory this is a single-facet model, and we are interested in the generalizability of scores across raters. Three variance components must be estimated: those due to essays, raters, and the rater-by-essay interaction. In the ideal situation, all papers would be read by all raters, and the estimation of these components would be rather routine. Since that was not the case in this study, it was necessary to take additional steps in order to obtain stable estimates. A more thorough treatment of generalizability theory in general,

and the procedures used to estimate variance components for this study in particular, may be found in Novak & Abedi (in preparation).

Percentages of agreement. Patterns of rater agreement differed between rubrics. While overall agreement for comparison ratings (Table 3) was generally satisfactory, it was lower and more variable across rater pairs than reliabilities achieved for previous studies (Gearhart, Herman, & Baker, 1992; Baker, Gearhart, & Herman, 1990, 1991). Rater agreement for the *Writing What You Read* ratings was generally acceptable, and somewhat higher and more consistent than that for comparison ratings. It was also, however, somewhat lower than the very high rates of agreement we have obtained for the comparison rubric in prior studies. There were no consistent differences among rater pairs in levels of agreement, nor any evident patterns among the subscales in levels of agreement.

While the agreements reported in these tables were certainly satisfactory, they were not exemplary. The patterns of rater agreement obtained here may have been impacted by study purpose: Raters were informed from the outset that they were participating in a study of two narrative rubrics, and they were atypically slow, methodical, and analytic in their approach to scoring, raising and pursuing issues that are often handled quickly and dismissed in moderation sessions. We suspect that moderation discussions confronted the raters with the complexity and uncertainty of the rating process.

Pearson correlations. The average correlations for the Overall, Character, and Communication scales for the WWYR rubric (Table 4) are quite comparable to those obtained for the three subscales for the Comparison rubric (Table 3), while those for the Theme, Setting, and Plot subscales were somewhat lower. The Plot scale was particularly problematic, with an average correlation of .48. Correlations across rater pairs were generally more consistent for the WWYR rubric, although this may be due largely to more stable estimates resulting from the larger number of papers that were scored using the WWYR rubric. Note that for the Comparison rubric the lowest correlations were obtained for the one and five pairing of raters (.28 and .25 for the General Competence and the Focus/Organization scales, respectively); those estimates, however, were based on a sample of only twelve papers.

Table 3

Rater Agreement: Comparison Rubric

Index and Raters	General Competence	Focus/Organization	Development/ Elaboration
Pearson correlation coefficients			
Raters 1 and 2 (N=18)	.51*	.51*	.63**
Raters 1 and 3 (N=21)	.56**	.57**	.39
Raters 1 and 4 (N=18)	.82**	.78**	.71**
Raters 1 and 5 (N=12)	.28	.25	.70*
Raters 2 and 3 (N=20)	.79**	.71**	.73**
Raters 2 and 4 (N=18)	.73**	.56*	.51*
Raters 2 and 5 (N=16)	.84**	.59**	.69**
Raters 3 and 4 (N=18)	.88**	.85**	.90**
Raters 3 and 5 (N=15)	.61*	.53*	.46
Raters 4 and 5 (N=19)	.73**	.62**	.61**
Average	**.68**	**.60**	**.63**
Percent agreement ±0			
Raters 1 and 2	.50	.50	.44
Raters 1 and 3	.38	.38	.24
Raters 1 and 4	.39	.39	.39
Raters 1 and 5	.33	.08	.42
Raters 2 and 3	.40	.15	.20
Raters 2 and 4	.28	.28	.33
Raters 2 and 5	.38	.13	.13
Raters 3 and 4	.56	.44	.39
Raters 3 and 5	.20	.20	.20
Raters 4 and 5	.32	.26	.32
Average	**.37**	**.28**	**.31**

* $p<.01$. ** $p<.05$.

Table 3 (continued)

Index and Raters	General Competence	Focus/Organization	Development/ Elaboration
Percent agreement ±0.5			
Raters 1 and 2	.83	.72	.72
Raters 1 and 3	.71	.57	.52
Raters 1 and 4	.78	.78	.72
Raters 1 and 5	.50	.50	.58
Raters 2 and 3	.70	.75	.50
Raters 2 and 4	.67	.61	.72
Raters 2 and 5	.81	.69	.75
Raters 3 and 4	78	.67	.89
Raters 3 and 5	.73	.40	.67
Raters 4 and 5	.79	.74	.63
Average	**.73**	**.64**	**.67**
Percent agreement ±1.0			
Raters 1 and 2	.94	.94	.94
Raters 1 and 3	.86	.81	.81
Raters 1 and 4	1.00	1.00	1.00
Raters 1 and 5	.67	.75	.83
Raters 2 and 3	1.00	.95	1.00
Raters 2 and 4	.89	.94	.89
Raters 2 and 5	1.00	1.00	1.00
Raters 3 and 4	1.00	.94	1.00
Raters 3 and 5	.87	.87	.93
Raters 4 and 5	1.00	1.00	1.00
Average	**.92**	**.92**	**.94**

* $p<.01$. ** $p<.05$.

Table 4

Rater Agreement: *Writing What You Read* Rubric

Level	Overall Effectiveness	Theme	Character	Setting	Plot	Communication
Pearson correlation coefficients						
Raters 1 and 2 (N=48)	.51**	.52**	.56**	.47**	.55**	.63**
Raters 1 and 3 (N=48)	.75**	.64**	.80**	.47**	.71**	.75**
Raters 1 and 4 (N=27)	.80**	.77**	.79**	.67	.71	.82**
Raters 1 and 5 (N=37)	.75**	.61**	.69**	.58**	.57**	.50**
Raters 2 and 3 (N=59)	.60**	.41**	.64**	.49**	.50**	.66**
Raters 2 and 4 (N=53)	.70**	.60**	.77**	.59	.66	.77
Raters 2 and 5 (N=58)	.61**	.52**	.61**	.16	.46**	.63**
Raters 3 and 4 (N=42)	.52**	.64**	.44**	.48**	.45**	.52**
Raters 3 and 5 (N=93)	.54**	.61**	.58**	.40**	.50**	.67**
Raters 4 and 5 (N=44)	.64**	.56**	.71**	.53**	.58**	.64**
Average	**.64**	**.59**	**.66**	**.48**	**.57**	**.66**
Percent agreement ±0						
Raters 1 and 2	.44	.38	.40	.44	.40	.42
Raters 1 and 3	.52	.46	.55	.50	.48	.42
Raters 1 and 4	.56	.52	.56	.56	.48	.56
Raters 1 and 5	.41	.46	.47	.51	.41	.32
Raters 2 and 3	.41	.29	.31	.34	.32	.51
Raters 2 and 4	.40	.43	.45	.53	.43	.47
Raters 2 and 5	.52	.34	.47	.31	.26	.47
Raters 3 and 4	.43	.38	.29	.31	.29	.38
Raters 3 and 5	.42	.42	.36	.46	.43	.44
Raters 4 and 5	.45	.45	.43	.55	.41	.39
Average	**.46**	**.41**	**.43**	**.45**	**.39**	**.44**

* $p<.01$. ** $p<.05$.

Table 4 (continued)

Level	Overall Effectiveness	Theme	Character	Setting	Plot	Communication
Percent agreement ±0.5						
Raters 1 and 2	.77	.75	.57	.73	.65	.75
Raters 1 and 3	.83	.69	.81	.75	.77	.79
Raters 1 and 4	.93	.74	.85	.81	.85	.89
Raters 1 and 5	.89	.59	.67	.68	.84	.70
Raters 2 and 3	.80	.73	.64	.63	.58	.85
Raters 2 and 4	.85	.77	.85	.79	.79	.89
Raters 2 and 5	.88	.72	.72	.53	.60	.86
Raters 3 and 4	.86	.79	.64	.71	.76	.81
Raters 3 and 5	.83	.73	.71	.72	.82	.84
Raters 4 and 5	.84	.68	.77	.77	.89	.82
Average	**.85**	**.72**	**.72**	**.71**	**.76**	**.82**
Percent agreement ±1.0						
Raters 1 and 2	.92	.90	.91	.85	.96	.96
Raters 1 and 3	.96	.96	.96	.94	.96	.96
Raters 1 and 4	1.00	1.00	1.00	1.00	1.00	1.00
Raters 1 and 5	.97	.92	.97	.97	.92	.97
Raters 2 and 3	.98	.93	.93	.92	.93	.97
Raters 2 and 4	1.00	.92	.98	.94	.96	.98
Raters 2 and 5	.95	.97	.86	.90	.95	.98
Raters 3 and 4	.93	.95	.90	.95	.95	.95
Raters 3 and 5	.96	.95	.97	.94	.95	.98
Raters 4 and 5	.95	.95	.98	.93	.95	.93
Average	**.96**	**.95**	**.95**	**.93**	**.95**	**.97**

* $p<.01$. **$p<.05$.

Generalizability coefficients. Table 5 shows the proportions of variance attributable to Essays, Raters, and to the Essay-by-Rater interaction, and the resultant generalizability coefficients. Coefficients for both relative and absolute decisions are reported. Note that for both rubrics the proportion of variance due to Raters is almost negligible. This indicates quite good consistency in the application of the scoring rubrics across raters, and has very positive implications with respect to the feasibility of using scores based on these rubrics to make absolute decisions about students' proficiencies, such as assignments to proficiency categories based on cutpoints, or comparisons of scores assigned to students by different raters. If the variance due to raters were large, then we would have very little assurance that scores assigned to students by different raters were based on the same scale. That is, if this were the case, then we could not be confident that a 3 given by one rater indicated the same level of proficiency as a 3 given by another rater. It is possible for raters to agree perfectly with respect to relative decisions and still not agree well with respect to absolute decisions. For example, if two raters scored a set of papers, and one rater always gave each paper a score that was 3 units higher than that awarded by the other rater, then the relative generalizability of those two raters would be perfect, while the absolute generalizability would be low. This is not the case here, however, and the very small variance components for raters ensure that the generalizability coefficients for relative and absolute decisions will be quite close together, as we see in Table 5.

Comparisons across rubrics. Comparing across rubrics and scales, we see that the G-coefficients for the Comparison rubric scales tend to be consistently higher than those for the WWYR rubric. G-coefficients for the Comparison rubric are quite consistent across scales, while there is considerable variation in the generalizability for the WWYR subscales, with the Setting subscale the most problematic with an estimated generalizability coefficient of 0.47.

D-study coefficients. If we compare the results in Table 5 with those in Tables 2 and 3, we see that the generalizability coefficients agree closely with the average Pearson correlations. The generalizability coefficients for relative decisions reported in Table 5 can be interpreted as reliability coefficients for scores based on a single rater, and those estimates are somewhat lower than would be desired. Although there are no cut-and-dried guidelines for what

Table 5

Generalizability Coefficients

		Variance components			*Generalizability coefficients*	
Rubric	Scale	E	R	ER	Relative	Absolute
Comparison	General Competence	0.68	0.00	0.32	0.68	0.68
	Focus/ Organization	0.63	0.01	0.36	0.64	0.63
	Development/ Elaboration	0.66	0.01	0.34	0.66	0.65
WWYR	Overall	0.60	0.01	0.40	0.60	0.59
	Theme	0.55	0.04	0.41	0.57	0.55
	Character	0.62	0.01	0.37	0.63	0.62
	Setting	0.47	0.00	0.53	0.47	0.47
	Plot	0.55	0.00	0.45	0.55	0.55
	Communication	0.62	0.00	0.37	0.63	0.63

Note. Standardized variance component estimates for Essay (E), Rater (R), and the Essay-by-Rater interaction (ER), and the generalizability coefficients derived from those estimates, for each of the Comparison and WWYR scales.

determines an adequate level of reliability, most researchers would probably like to see reliabilities of at least .75. The generalizability coefficients for both rubrics fall well below that threshold. The next step within the context of generalizability theory was to use the results of the G-study to perform a D-study in order to determine how to attain an acceptable reliability level. Table 6 reports D-study generalizability coefficients for scores based on 1, 2, 3, and 5 raters.

Table 6

D-study Coefficients

		Relative				*Absolute*			
Rubric	Scale	1	2	3	5	1	2	3	5
Comparison	General Competence	0.68	0.81	0.86	0.91	0.68	0.81	0.86	0.91
	Focus/ Organization	0.64	0.78	0.84	0.90	0.63	0.77	0.84	0.89
	Development/ Elaboration	0.66	0.80	0.85	0.91	0.65	0.79	0.85	0.90
WWYR	Overall	0.60	0.75	0.82	0.88	0.59	0.75	0.81	0.88
	Theme	0.57	0.73	0.80	0.87	0.55	0.71	0.79	0.86
	Character	0.63	0.77	0.83	0.89	0.62	0.77	0.83	0.89
	Setting	0.47	0.64	0.73	0.82	0.47	0.64	0.73	0.82
	Plot	0.55	0.71	0.79	0.86	0.55	0.71	0.79	0.86
	Communication	0.63	0.77	0.83	0.89	0.63	0.77	0.83	0.89

Note. D-study generalizability coefficients for relative and absolute decisions for essay scores based on 1, 2, 3, or 5 raters.

The results of the D-study show that for all of the Comparison subscales and for three of the WWYR subscales, adequate reliability (as defined above) can be obtained through the use of two raters. Note, however, that for the WWYR Setting subscale, even the use of three raters is not sufficient to ensure a reliability level of .75. Using four raters would result in a coefficient of .78 for this scale. Again, due to the very small proportions of variance attributable to the Rater main effects, results and interpretations for relative and absolute decisions are nearly identical.

Validity

Validity of the Writing What You Read *rubric: What is the evidence that scores derived from the WWYR rubric are meaningful indices of students' narrative writing?*

This section contains four analyses of the *Writing What You Read* rubric's capacity to produce meaningful results: (a) comparisons of students' scores across grade levels (scores should increase with grade level); (b) intercorrelations of subscales within rubrics (for each rubric, subscales should not be highly correlated); (c) correlations of ratings across rubrics (WWYR scores should correlate significantly with comparison scores); (d) an analysis of decision consistency across rubrics (raters should make similar decisions about students' competence across rubrics). All ratings contributed to these results: Paper scores were computed as the average of the independent ratings *or* the resolved score achieved through discussion during the training and check sets.

Grade level comparisons. Tables 7 and 8 contain descriptive statistics for each rubric and, for each subscale, the results of ANOVAs by Level. For each rubric, there were score differences in the expected direction by grade level. The pattern of score differences was the same for all scales and both rubrics, although the ANOVA result for one WWYR subscale (Plot) was not significant.

Intercorrelations of subscales within rubrics. Tables 9 and 10 contain intercorrelations of subscales for each rubric. All subscales were highly correlated, indicating that raters were not making highly differentiated judgments about a narrative's competence along each dimension. Based on these results, subscales for both rubrics are not empirically distinct.

Correlations of ratings across rubrics. Table 11 contains intercorrelations of subscales for each rubric. Across rubrics, scores were highly intercorrelated, although the correlations were lower in magnitude than the within-rubric correlations (Tables 9 and 10).

Table 7

Descriptives, Comparison Rubric

Level	Subscale		
	General Competence	Focus/ Organization	Development/ Elaboration
Primary (*N*=16)			
Mean	2.05	2.29	2.27
SD	.47	.48	.45
Middle (*N*=36)			
Mean	2.58	2.68	2.79
SD	.55	.50	.59
Upper (*N*=17)			
Mean	3.54	3.66	3.67
SD	.49	.67	.57

Note. For this analysis, *N* = number of subjects. ANOVAs examining differences among Levels for each scale: General Competence, F (2,66) = 36.380, $p < .0001$; Focus/Organization, F (2,66) = 29.136, $p < .0001$; Development/Elaboration, F (2,66) = 26.978, $p < .0001$.

Table 8

Descriptives, *Writing What You Read* Rubric

Level	Subscale					
	Overall	Theme	Character	Setting	Plot	Commun-ication
Primary (*N*=17)						
Mean	2.29	2.47	2.15	2.27	2.44	2.33
SD	.39	.48	.53	.42	.49	.44
Middle (*N*=36)						
Mean	2.50	2.61	2.40	2.49	2.55	2.51
SD	.44	.45	.53	.43	.47	.49
Upper (*N*=20)						
Mean	2.87	3.02	2.78	2.73	2.80	2.96
SD	.59	.64	.74	.51	.64	.64

Note. For this analysis, *N* = number of subjects. ANOVAs examining differences among Levels for each scale: Overall, $F(2,70) = 7.113$, $p < .002$; Theme, $F(2,70) = 6.105$, $p < .004$; Character, $F(2,70) = 5.445$, $p < .006$; Setting, $F(2,70) = 4.929$, $p < .01$; Plot, $F(2,70) = 2.473$, $p < .092$; Communication, $F(2,70) = 7.519$, $p < .001$.

Table 9

Subscale Correlations, Comparison Rubric (N=184)

Level and Subscale	Subscale: General Competence	Subscale: Focus/ Organization	Subscale: Development/ Elaboration
Primary (N=36)			
General Competence		.80*	.81*
Focus/Organization			.74*
Middle (N=115)			
General Competence		.87*	.90*
Focus/Organization			.80*
Upper (N=35)			
General Competence		.91*	.86*
Focus/Organization			.82*
Overall (N=184)			
General Competence		.91*	.92*
Focus/Organization			.85*

*$p < .001$.

Table 10

Subscale Correlations, *Writing What You Read* Rubric (N=187)

	Scale					
Scale	Overall	Theme	Character	Setting	Plot	Communication
Primary (N=37)						
Overall		.88*	.86*	.86*	.86*	.86*
Theme			.85*	.73*	.87*	.83*
Character				.77*	.82*	.81*
Setting					.77*	.82*
Plot						.82*
Middle (N=112)						
Overall		.92*	.91*	.87*	.93*	.94*
Theme			.88*	.81*	.89*	.89*
Character				.85*	.88*	.88*
Setting					.82*	.81*
Plot						.92*
Upper (N=38)						
Overall		.94*	.90*	.92*	.95*	.97*
Theme			.90*	.91*	.93*	.95*
Character				.83*	.91*	.91*
Setting					.89*	.92*
Plot						.94*
Total (N=187)						
Overall		.93*	.91*	.89*	.93*	.94*
Theme			.90*	.84*	.90*	.90*
Character				.84*	.89*	.89*
Setting					.84*	.85*
Plot						.91*

*$p < .001$.

Table 11

Correlations Across Rubrics

Comparison Scale	*Writing What You Read* Scale					
	Overall	Theme	Character	Setting	Plot	Communication
Primary(N=36)						
General Competence	.62***	.61***	.70***	.59***	.69***	.68***
Focus/ Organization	.46*	.54**	.44*	.43*	.56***	.58***
Development/ Elaboration	.62***	.60***	.61***	.65***	.65***	.64***
Middle (N=107)						
General Competence	.79***	.75***	.75***	.71***	.72***	.77***
Focus/ Organization	.71***	.68***	.65***	.60***	.66***	.68***
Development/ Elaboration	.74***	.71***	.70***	.65***	.70***	.74***
Upper (N=33)						
General Competence	.74***	.71***	.72***	.68***	.74***	.73***
Focus/ Organization	.65***	.55***	.59***	.56***	.65***	.64***
Development/ Elaboration	.67***	.62***	.71***	.60***	.65***	.68***
Total (N=176)						
General Competence	.75**	.73**	.74**	.66**	.67**	.74**
Focus/ Organization	.67**	.66**	.64**	.58**	.62**	.68**
Development/ Elaboration	.72**	.70**	.71**	.64**	.66**	.73**

*$p < .05$. **$p < .01$. ***$p < .001$.

Decision consistency across rubrics. To examine consistency in raters' judgments of narrative competence across rubrics, we cross-classified scores for General Competence (comparison) and Overall Effectiveness (WWYR) (Table 12). These results must be interpreted in the context of two important issues. First, although both rubrics are 6-point scales, their scale points do not correspond in meaning; in particular, the WWYR rubric is developmental and is not intended to locate competency at any particular level. Second, although the "best fit" for WWYR's definition of a competent narrative may be Level 3 ("One episode narrative (either brief or more extended) which includes the four critical elements of problem, emotional response, action, and outcome. . . . "), the criteria for this level were considered unclear by our raters, as we discuss below.

We chose a WWYR mean rating of 3 or above as evidence of competence, and compared WWYR judgments against comparison ratings of 3.5 or above, consistent with the comparison rubric's distinction between a "developing writer" (Level 3) and a "competent writer" (Level 4). Most papers were judged as lacking in competence. Raters agreed in their classifications of 146 of 176 papers (Pearson, $p<.00001$). However, there was no consistent agreement in classification of "competent" papers: Of the 55 papers judged as competent with either rubric, only 25 were classified as competent with both rubrics.

Table 12

Cross-Classification of Comparison and WWYR Scores (N=176)

	Comparison General Competence	
WWYR Overall Effectiveness	<3.0	= or > 3.0
< 2.5	121	14
= or > 2.5	16	25

Note. For each rubric, each paper was scored by at least two raters; paper scores were computed as the mean of all raters' judgments.

Raters' Reflections

What are raters' views of the utility and validity of the comparison and Writing What You Read *rubrics?*

Raters were interviewed at two points in the rating process—following comparison ratings (a focus group discussion) and following completion of ratings with both rubrics (an interview with pairs of raters). At each interview, raters scored sample narratives and discussed the fit of the rubrics to the papers. The results reported below highlight the raters' comparisons of the WWYR to the comparison rubric. Raters raised concerns regarding rubric content, ease of use, instructional potential, and feasibility for large-scale scoring.

Rubric content. Raters offered a balanced appraisal of the strengths of each rubric. Raters viewed WWYR as more comprehensive in its analysis of narrative, more "positive" in each of its scale-point definitions (more specific about narrative qualities and less "negative" or comparative), and more complete in its analysis of a narrative's "development." The content "missing" in the comparison Development/Elaboration subscale was first discussed even prior to the raters' introduction to WWYR, when raters explained that they had added content that they considered central to their judgment of narrative: "I put feeling under Elaboration. I know it's not, but . . . you need to." "There's a big difference between actually seeing something visually and feeling something . . . [S]omething can be 'vivid,' and something can be 'elaborate,' but it might not make you feel emotionally."

In their critique of WWYR content, raters focused on Plot, Overall Effectiveness, Communication, and the absence of a scale like comparison's Focus/Organization. Plot and Overall Effectiveness were seen as weak in their middle sections, handling ineffectively those narratives that contained a series of incomplete episodes. Communication was considered helpful in pinpointing particular techniques, but its emphasis on language choices "appropriate to the narrative" made it difficult for the raters to give a child credit for stylistic strength that did not necessarily contribute to the narrative. In addition, they felt that Communication could be differentiated—at least for

instructional applications—as separate subscales for style, tone, and voice.[3] Finally, raters missed using comparison's Focus/Organization scale. While this comparison scale was seen as rather dry and perhaps exposition-like, it captured for these raters a dimension of organizational competence missing in WWYR.

Raters felt that neither rubric was able to capture a narrative's local strengths: "Maybe they have one character description, or a setting, or something funny, and you laugh, but it really doesn't allow itself to be 4 and you want to tell them, 'Hey, you made me laugh here, or look at all these similes you were using.' " Similarly, some raters felt that neither rubric represented creativity very well: "There might be some idiosyncratic quality or some uniqueness about it, some originality that you can't really score." Wanting to "give credit" to a child for a moment of insight, humor, language use, or cleverness, they suggested providing a place on the rating form for personal comments to each writer on strengths and weaknesses.

Ease of use. Although most raters felt that application of the WWYR rubric was a slower, more "analytical" process than comparison rating, only one of the five raters remained uncomfortable: "[The WWYR rubric is] so broken apart, analytic, that it confuses me." Indeed, the WWYR rubric did contain a greater number of scales and detail at each scale point, and, for this rater, the constructs required explication ("explicit and implicit, didactic and revealing . . . it's too much to keep track of"). For the remaining raters, the acknowledged difficulty of WWYR scoring was balanced with enjoyment.

> It was much easier, much more enjoyable to use the WWYR to score it. Because [the rubric] talked about the different subtleties of language and the different styles and emotions that you could use to make it more sophisticated and improve it. Whereas the comparison didn't really give that feeling . . . [L]anguage . . . just seemed like a skill rather than a quality of the work.

Raters also appreciated the specificity of the WWYR rubric. Four of the five raters reported difficulty anchoring comparison judgments based on comparative criteria: "This 'few, many, little, and more' kind of vocabulary . . . was really a problem in the beginning . . . What is 'many?' What is 'few?'

[3] An early version of the WWYR rubric in fact contained these dimensions. Copies of the rubric draft are available from the authors.

We had to make our own kind of interpretations, and then compare as we went on reading." Wishing for more positive and specific descriptions, one rater commented: "What is the paper *do*ing, even though there might be inappropriate [language]. . . . 'No development of narrative elements'—what can you say instead of that?" To adapt, raters reported several strategies for resolving uncertainty: expanding the list of comparison criteria (the addition of "emotion" to Development, as discussed above); making iterative comparisons with higher and lower scale points; using the anchor terms in the left column; making an initial dichotomous judgment between "Developing" (1-3) and "Competent" (4-6) writer and then refining the decision. WWYR, in contrast, supported greater focus on the fit of a narrative to the characteristics listed at a given level.

The raters' response to WWYR was encouraging. Their relative comfort indicated that a two- to three-hour WWYR training session can be adequate for many raters, if they are experienced with scoring and knowledgeable about narrative. Raters did offer suggestions for improvements of the WWYR rubric that would have facilitated scoring for them: highlighting key terms, listing criteria as bullets, and adopting overarching descriptors like those in comparison's left column (e.g., Developing Writer, Competent Writer).

Instructional potential. Most raters viewed the WWYR rubric as having far more instructional potential than the comparison rubric, and those four raters who were classroom teachers planned to utilize it in some form in their classrooms. For example, one of the comparison Valley raters commented:

> [WWYR] allows you to compliment other strengths, and their styles . . . It's wonderful to have it for a teacher resource to direct the children, and the parents . . . When I'm scoring kids [with comparison], I'm having a hard time putting into words what I want them to do. With WWYR, I could get up and directly teach a lesson.

But one of the four teachers felt that WWYR demanded more analysis than she could routinely or profitably undertake in the classroom. For this rater, difficulty of use limited instructional potential: "For many teachers, you have to give them something that's easy to apply, an easy tool that we can use. . . . Not too much analyzing, not too much re-reading. Something automatic. I would like a tool like that . . . for our daily writing." A rubric with content as

complex as WWYR would be useful, she granted, when undertaking "a major project, then I want to use something like the *Writing What You Read,* if I want to touch on every single part [of the writing]."

Feasibility of use for large-scale assessment. Raters agreed that the comparison rubric had the capacity to be used reliably and with reasonable speed. In contrast, the feasibility and utility of WWYR for large-scale assessment were left as unanswered questions. The WWYR Overall Effectiveness scale was considered as a possible holistic replacement for comparison's General Competence, but there were concerns about the relation between the two judgments: Overall Effectiveness required a rater to judge the narrative's integration of other narrative elements, still a fairly analytic task that felt different in content and in process from a General Competence decision. Although raters acknowledged that they themselves had acquired expertise with WWYR in half a day, they nevertheless expressed concern about the staff development that would be required to implement a large-scale program based on WWYR assessment.

Summary and Discussion

The purpose of this study was to gather evidence of validity for a new narrative rubric designed to enhance the instructional value of writing assessments, but whose technical quality is unknown. The design of the *Writing What You Read* narrative rubric was prompted by the need for assessment tools that can guide instruction. The rubric differs from most narrative rubrics in its narrative-specific content and its developmental framework. Designed for classroom use and shown to impact elementary teachers' understandings of narrative (Gearhart et al., 1994), the rubric contains five analytic subscales for Theme, Character, Setting, Plot, and Communication, and a sixth, holistic scale for Overall Effectiveness. Each subscale contains six levels designed to match current understandings of children's narrative development. It has never been used to date for large-scale assessment.

Our study evaluated the *Writing What You Read* rubric against an established rubric that has consistently demonstrated sound technical capabilities in large-scale use. Our findings regarding the reliability and

validity of both rubrics yielded promising but mixed evidence of the utility of the *Writing What You Read* rubric for large-scale assessment.

In general, both rubrics were used consistently by raters when making judgments of elementary children's classroom narratives. Rater agreements for three of the *Writing What You Read* subscales (Overall Effectiveness, Character, Communication) were consistent with those obtained with the comparison rubric, while levels of agreement for the other three WWYR subscales (Theme, Setting, Plot) were somewhat lower. Although overall agreement for both sets of ratings was generally satisfactory, it was lower and more variable across rater pairs than reliabilities achieved for previous studies; the WWYR rubric did not exhibit as much variation across rater pairs as did the comparison rubric. Results of the generalizability analyses indicated that adequate levels of reliability for most scales of either rubric could be attained by doubly scoring each essay and aggregating the results. However, for WWYR, achieving adequate reliability for Setting (and, to a lesser degree, Theme and Plot) could require as many as four raters.

The patterns of rater agreement obtained here may have been impacted by both study purpose and rubric content. First, raters were informed from the outset that they were participating in a study of two narrative rubrics, and they were atypically slow, methodical, and analytic in their approach to scoring, raising and pursuing issues that are often handled quickly and dismissed in moderation sessions. Second, the WWYR rubric's representations of certain aspects of narrative competence were issues from the beginning of WWYR scoring. Although findings for raters' comments and the quantitative analyses were consistent only for Plot (in that both data sources pointed to content weaknesses), the overall findings do indicate a need to revisit aspects of the content of the rubric.

There were several sources of evidence for the validity of the *Writing What You Read* rubric. First, the scores from both rubrics produced a pattern of increasing competence with grade level. Second, WWYR scores were highly correlated with the comparison scores, although there was some evidence for the distinctiveness of the two scales in the finding that cross-rubric subscale correlations were lower than within-rubric subscale correlations. Third, comparisons of raters' judgments made with both rubrics for the same narratives indicated some consistency in their decisions, although

disagreements in classifications of "competent" narratives suggested distinctive definitions for competence. Finally, raters felt that the content of the WWYR rubric captured more aspects of narrative than the comparison rubric and had greater instructional potential. However, raters perceived some distinctive utility in the comparison Focus/Organization scale, and they recommended revisions of the scales for Plot, Overall Effectiveness, and Communication. They also expressed some concern about the professional development that would be required for WWYR scoring, despite their recognition that they had achieved understandings of the WWYR rubric and consensus in its use after only a two-hour training session.

Thus our study has produced evidence that at least three scales of the *Writing What You Read* narrative rubric—an analytic writing rubric designed to enhance teachers' understandings of narrative and to inform instruction—can be used reliably and meaningfully in large-scale assessment of elementary level writing, provided that each narrative is rated by two raters. While we would have preferred that our analyses yield evidence of the technical soundness of all six scales, it is nevertheless heartening that a rubric as substantive as WWYR could produce findings this positive in an initial study.

An important issue remains unresolved. Consistent with other studies of analytic scales, neither the WWYR nor the comparison rubric produced patterns of highly distinctive subscale judgments. We produced no empirical evidence for the subscales of either rubric. While raters agreed that WWYR scales had greater instructional utility than comparison scales and that each of the WWYR scales had relevance for instructional planning and classroom assessment, our quantitative findings suggest that subscale judgments may not provide a technically sound profile of students' strengths and weaknesses.

We do not view these findings as a basis for rejecting an analytic *framework* for scoring, although the results may have implications for the value of subscale scores. Further research is needed to determine the factors that support or constrain distinctive subscale judgments—the structure and content of analytic rubrics, the types of material to be rated, and the methods of rater training. If technical studies continue to demonstrate that subscale judgments can not be distinguished from overall competence ratings, we would argue for some "analytic" alternative to holistic scoring. One option

might be assignment of a single score, supplemented with rater commentary on strengths and weaknesses guided by checklists or open-ended prompts.

Writing rubrics represent frameworks for interpretation of text and have potential to enhance teachers' knowledge and practice. When rubrics are designed to capture qualities of distinctive writing genres, then they have greater potential to support teachers' professional development, opportunities to learn in the classroom, and substantive interactions in moderation sessions.

References

Baker, E.L., Gearhart, M., & Herman, J.L. (1990). *The Apple Classrooms of Tomorrow: 1989 Evaluation Study* (Report to Apple Computer, Inc.). Los Angeles: University of California, Center for the Study of Evaluation.

Baker, E.L., Gearhart, M., & Herman, J.L. (1991). *The Apple Classrooms of Tomorrow: 1990 Evaluation Study* (Report to Apple Computer, Inc.). Los Angeles: University of California, Center for the Study of Evaluation.

Baker, E.L., Herman, J.L., & Gearhart, M. (1988). *The Apple Classrooms of Tomorrow: 1988 Evaluation Study* (Report to Apple Computer, Inc.). Los Angeles: University of California, Center for the Study of Evaluation.

Baxter, G.P., Glaser, R., & Raghavan, K. (in press). *Analysis of cognitive demand in selected alternative science assessments* (CSE Tech. Rep.). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Freedman, S. (May, 1991). Evaluating writing: *Linking large-scale assessment testing and classroom assessment* (Occasional Paper No. 27). Berkeley: University of California, Center for the Study of Writing.

Gearhart, M., Bank, A. & Herman, J.L. (1990). *Belridge DACOTT 21/20: The UCLA 1989-90 Evaluation* (Report to Belridge School). Los Angeles: University of California, Center for the Study of Evaluation.

Gearhart, M., Herman, J.L., & Baker, E.L. (1992, April) Writing portfolios at the elementary level: A study of methods for writing assessment. In L. Winters (Chair) *Perspectives on the assessment value of portfolios: How robust are they?* Symposium conducted at the annual meeting of the American Educational Research Association, San Francisco.

Gearhart, M., Herman, J.L., Baker, E.L., & Whittaker, A.K. (1992). *Writing portfolios at the elementary level: A study of methods for writing assessment* (CSE Tech. Rep. No. 337). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Gearhart, M., Herman, J.L., & Bank, A. (1989). *Belridge DACOTT 21/20: The UCLA 1988-89 Evaluation* (Report to Belridge School). Los Angeles: University of California, Center for the Study of Evaluation.

Gearhart, M., Wolf, S.A., Burkey, B., & Whittaker A.K. (1994). *Engaging teachers in assessment of their students' narrative writing: Impact on teachers' knowledge and practice* (CSE Tech. Rep. No. 377). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Herman, J.L., Baker, E.L., & Gearhart M. (1988). *The DACOTT 21/20 Belridge Evaluation Study* (Report to Belridge School). Los Angeles: University of California, Center for the Study of Evaluation.

Herman, J.L., Gearhart, M., & Baker, E.L. (1994). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment*, *1*(3), 201-224.

Huot, B. (1991). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*, 237-263.

LeMahieu, P., Gitomer, D.H., & Eresh, J.T. (in press). Portfolios in large-scale assessment: Difficult but not impossible.

Novak, J., & Abedi, J. (in preparation). *Estimation of variance components for incomplete crossed designs.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Paul, R.W. (1993). *Pseudo critical thinking in the educational establishment: A case study in educational malpractice.* Sonoma, CA: Sonoma State University, Center for Critical Thinking and Moral Critique.

Quellmalz, E., & Burry, J. (1983). *Analytic scales for assessing students' expository and narrative writing skills* (CSE Resource Paper No. 5). Los Angeles: University of California, Center for the Study of Evaluation.

Resnick, L., Resnick, D., & DeStefano, L. (1993). *Cross-scorer and cross-method comparability and distribution of judgments of student math, reading, and writing performance: Results from the New Standards Project Big Sky scoring conference* (CSE Tech. Rep. No. 368). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Wiggins, G. (1993, November). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, *75*(3), 200-208, 210-214.

Wolf, D.P. (1993). Assessment as an episode of learning. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 213 - 240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wolf, D.P., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education*, *17*, 31-74.

Wolf, S.A., & Gearhart, M. (1993a). *Writing What You Read: Assessment as a learning event.* (CSE Tech. Rep. No. 358). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Wolf, S.A., & Gearhart, M. (1993b). *Writing what you read. A guidebook for the assessment of children's narratives* (CSE Resource Paper No. 10). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Appendix

Interview Protocols